UNIFORMLY MOST POWERFUL TESTS FOR SIMULTANEOUSLY DETECTING A TREATMENT EFFECT IN THE OVERALL POPULATION AND AT LEAST ONE SUBPOPULATION

BY MICHAEL ROSENBLUM

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

After conducting a randomized trial, it is often of interest to determine treatment effects in the overall study population, as well as in certain subpopulations. These subpopulations could be defined by a risk factor or biomarker measured at baseline. We focus on situations where the overall population is partitioned into two predefined subpopulations. When the true average treatment effect for the overall population is positive, it logically follows that it must be positive for at least one subpopulation. We construct new multiple testing procedures that are uniformly most powerful for simultaneously rejecting the overall population null hypothesis and at least one subpopulation null hypothesis, when outcomes are normally distributed. We prove our procedures do not require any sacrifice for detecting a treatment effect in the overall population, compared to the uniformly most powerful test of the overall population null hypothesis. The proofs rely on a general method for transforming analytically difficult expressions arising in some multiple testing problems into more tractable nonlinear optimization problems, which are then solved using intensive computation.

1. Introduction. Planning a randomized trial of an experimental treatment can be challenging when it is suspected that certain populations may benefit more than others. For example, consider studies of metastatic breast cancer in which human epidermal growth factor receptor-2 (HER2) is overexpressed. The estimated benefit of trastuzumab treatment was greater for patients with higher levels of pretreatment HER2 overexpression (Slamon et al., 2001). As another example, a metastudy by Kirsch et al. (2008) of certain antidepressant medications suggests that there may be a clinically meaningful benefit only for those with severe depression at baseline.

We take the perspective of a researcher designing a randomized trial of a new treatment, where it is suspected that the magnitudes of treatment effects may differ in predefined subpopulations. We focus on the case of two predefined subpopulations that partition the overall study population, though we prove a result for $k > 2$ subpopulations as well.

AMS 2000 subject classifications: Primary 62F03; secondary 62F05, 62C20.

Keywords and phrases: Familywise Type I error, Maximin power, Nonlinear optimization, Personalized medicine

For a given population, the mean treatment effect is defined as the difference between the mean outcome were everyone assigned to the treatment and the mean outcome were everyone assigned to the control. We develop procedures to simultaneously test the null hypotheses of no positive mean treatment effect for subpopulation one (H_{01}), for subpopulation two (H_{02}), and for the overall study population (H_{0*}). These hypotheses are defined formally in Section 3.3 below. For each of these null hypotheses, the alternative hypothesis is that there is a positive mean treatment effect for the corresponding population.

In some cases, there is a subpopulation for which a larger treatment benefit is conjectured. We call this the favored subpopulation, and refer to the other as the complementary subpopulation. Since it is usually not known with certainty before the trial that the treatment will benefit the favored subpopulation, preplanning a hypothesis test for it is important. Also, in trials where the overall population null hypothesis is rejected, it is of clinical importance to determine if the treatment benefits the complementary subpopulation, since there was more a priori uncertainty about treatment effects for this group. Therefore, preplanning a hypothesis test for this subpopulation is also valuable. This motivates our interest in testing both subpopulation null hypotheses.

Our goal is to maximize power for simultaneously rejecting the overall population null hypothesis and at least one subpopulation null hypothesis. We give new multiple testing procedures that maximize this power, uniformly over all possible alternatives, in the case of two subpopulations and outcomes that are normally distributed. These procedures, which we denote by M^{UMP} and $M^{\text{UMP+}}$, are defined in Section 4.1. They require no sacrifice in detecting treatment effects for the overall population; that is, their probability of rejecting the null hypothesis H_{0*} equals that of the uniformly most powerful test of H_{0*}, for any data generating distribution.

In Section 5, we show our new procedures are consonant. According to Bittman et al. (2009) "A testing method is consonant when the rejection of an intersection hypothesis implies the rejection of at least one of its component hypotheses." Consonance was introduced by Gabriel (1969), and subsequent work on consonant procedures includes (Hommel, 1986; Romano and Wolf, 2005; Bittman et al., 2009; Brannath and Bretz, 2010; Romano, Shaikh and Wolf, 2011). Consonance is desirable since whenever an intersection of null hypotheses is false, it follows logically that at least one of the corresponding individual null hypotheses must be false as well. A non-consonant procedure is one that may reject an intersection of null hypotheses without rejecting any of the corresponding individual null hypotheses. For example, in our context a non-consonant procedure would sometimes make claims that logically imply the treatment is superior to control in at least one of the two subpopulations, without indicating which one. To the best of our knowledge, our procedures are the first multiple testing procedures for our problem that are

consonant.

The above properties hold for our procedures regardless of the relative sizes of the two subpopulations. Though we focus on one-sided tests, we also construct a consonant procedure corresponding to two-sided tests, and prove it satisfies a maximin optimality property.

We prove two impossibility results, showing certain properties cannot be simultaneously satisfied by any multiple testing procedure. First, consider any data generating distribution Q for which H_{0*} is false. This implies that under Q, at least one subpopulation null hypothesis is false. Assume outcomes under treatment and control, for each subpopulation, are normally distributed under Q. We show no multiple testing procedure has all of the following properties:

i. Equal or greater power at Q for simultaneously rejecting the overall population null hypothesis and at least one subpopulation null hypothesis, compared to our procedure M^{UMP}.
ii. It dominates a multiple testing procedure of (Rosenbaum, 2008, Section 2), which we describe in Section 6 below.
iii. Strong control of the familywise Type I error rate at level 0.05.

Though the procedure of (Rosenbaum, 2008, Section 2) neither satisfies (i) nor is consonant, it does have important advantages described below. Our second negative result is that for more than two subpopulations, it is not possible simultaneously to be consonant (as defined in Section 5) and to have probability of rejecting H_{0*} at least that of the uniformly most powerful test of H_{0*}.

2. Related work. Multiple testing procedures for the family of null hypotheses H_{0*}, H_{01}, H_{02} can be constructed using the methods of, e.g., Holm (1979), Bergmann and Hommel (1988), Maurer, Hothorn and Lehmacher (1995), Song and Chi (2007), Rosenbaum (2008), and Alosh and Huque (2009). Each of these strongly controls the familywise Type I error rate, as defined by Hochberg and Tamhane (1987). However, none of these procedures is uniformly most powerful for simultaneously rejecting the overall population null hypothesis and at least one subpopulation null hypothesis, as defined in Section 3.6. We compare the power of our uniformly most powerful procedures to this prior work, in Section 7.

3. Multiple testing problem.

3.1. *Randomized trial description.* We consider two-armed randomized trials. Membership in each subpopulation must be a prespecified function of prerandomization variables. For each subpopulation $s \in \{1, 2\}$, let $p_s > 0$ denote the fraction of the overall population in subpopulation s. Since by assumption the subpopulations partition the overall population, we have $p_1 + p_2 = 1$.

Let $a = 1$ denote the treatment arm and $a = 0$ denote the control arm. We let n_{sa} denote the number of participants in subpopulation $s \in \{1, 2\}$ who are assigned to study arm $a \in \{0, 1\}$. Denote the total sample size by n. We assume the fraction of trial participants in each subpopulation s, $(n_{s0} + n_{s1})/n$, equals the corresponding population proportion p_s. We also assume the proportion of participants from each subpopulation that are assigned to the treatment arm is $1/2$; this can be approximately ensured if block randomization is used for each subpopulation. We conjecture our results extend to the case of unequal randomization probabilities, but this is an area for future research.

3.2. *Data collected on each participant.* For each participant i, denote his/her subpopulation by $S_i \in \{1, 2\}$, study arm assignment by $A_i \in \{0, 1\}$, and outcome by $Y_i \in \mathbb{R}$. We assume for each participant i, conditioned on his/her subpopulation $S_i = s$ and study arm assignment $A_i = a$, that his/her outcome Y_i is a random draw from an unknown distribution Q_{sa} and this draw is independent of the data of all other participants. Let $\mu(Q_{sa})$ denote the mean and $\sigma^2(Q_{sa})$ denote the variance of the outcome distribution Q_{sa} for subpopulation $s \in \{1, 2\}$ and study arm $a \in \{0, 1\}$. For compactness we represent $(Q_{10}, Q_{11}, Q_{20}, Q_{21})$ by Q.

We make no parametric model assumptions on the form of each outcome distribution Q_{sa}. Instead, we make a weaker assumption, given next. For fixed $C > 0$, let $\mathcal{Q}$ denote the class of data generating distributions Q whose components Q_{sa} each satisfy

$$E_{Q_{sa}} |Y - \mu(Q_{sa})|^3 / \{\sigma^2(Q_{sa})\}^{3/2} < C. \tag{3.1}$$

This condition, combined with the multivariate, Berry-Esseen central limit theorem of Götze (1991), implies the joint distribution of subpopulation-specific z-statistics (defined below) converges uniformly to a multivariate normal distribution. Such uniform convergence is generally required to ensure that even the standard, one-sided z-test strongly controls the asymptotic, familywise Type I error rate, in the uniform sense that we define in the next section.

Condition (3.1) is satisfied, for example, by any Q whose components Q_{sa} are normally distributed, as long as we set $C > 2$. Also, for fixed $K > 0$ and $\tau > 0$, condition (3.1) is satisfied for the class of distributions where each Q_{sa} has support in $[-K, K]$ and variance at least τ, if we set $C > (2K)^3/\tau^{3/2}$. We assume C is a fixed value, i.e., it does not depend on sample size.

3.3. *Hypotheses tested.* The null hypotheses to be tested, which correspond to no positive mean treatment effect in subpopulation one, in subpopulation two, and

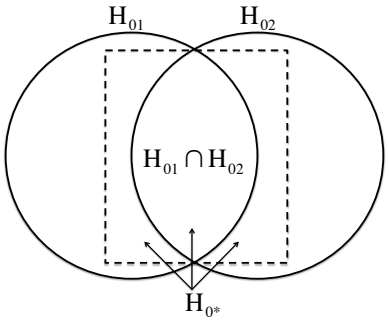

FIG 1. *The relationship among the null hypotheses H_{01}, H_{02}, and H_{0*}. The dashed rectangle corresponds to the null hypothesis H_{0*} for the overall population.*

in the overall population, respectively, are

$$H_{01} = \{Q \in \mathcal{Q} : \mu(Q_{11}) - \mu(Q_{10}) \leq 0\}; \tag{3.2}$$

$$H_{02} = \{Q \in \mathcal{Q} : \mu(Q_{21}) - \mu(Q_{20}) \leq 0\}; \tag{3.3}$$

$$H_{0*} = \{Q \in \mathcal{Q} : p_1 \{\mu(Q_{11}) - \mu(Q_{10})\} + p_2 \{\mu(Q_{21}) - \mu(Q_{20})\} \leq 0\}. \tag{3.4}$$

We refer to these as elementary null hypotheses. The corresponding alternative hypotheses are the complements of each of these null hypotheses. We prove results for the null hypotheses of zero mean treatment effect, and the corresponding two-sided alternative hypotheses, in Section 4.2.

The key relationship among the above null hypotheses is:

$$(H_{01} \cap H_{02}) \subset H_{0*} \subset (H_{01} \cup H_{02}). \tag{3.5}$$

The set of distinct intersections of the three elementary null hypotheses is

$$\mathcal{H} = \{H_{0*}, H_{01}, H_{02}, H_{01} \cap H_{02}, H_{0*} \cap H_{01}, H_{0*} \cap H_{02}\}.$$

The intersection $H_{01} \cap H_{02} \cap H_{0*}$ does not appear in the above list since it equals $H_{01} \cap H_{02}$.

We focus on the set of null hypotheses $\{H_{01}, H_{02}, H_{0*}\}$ rather than the simpler set $\{H_{01}, H_{02}, H_{01} \cap H_{02}\}$. The reason is that the primary hypothesis in many randomized trials concerns the average effect of treatment versus control in the overall population, as represented, e.g., in H_{0*}. If H_{0*} is false, the clinical implication is that giving everyone in the overall population the treatment, rather than giving everyone the control, improves the average outcome. In contrast, if the intersection

hypothesis $H_{01} \cap H_{02}$ is false, the aforementioned clinical implication does not necessarily hold.

A multiple testing procedure M is defined as a deterministic map from the data generated in the randomized trial described in Sections 3.1 and 3.2 to a subset of $\{H_{01}, H_{02}, H_{0*}\}$ representing the null hypotheses that are rejected. In order that probabilities such as the familywise Type I error rate of M are well-defined, we assume each M satisfies a measurability condition that we define next.

For total sample size n, denote the class of possible data sets by $\Omega = (\{1,2\} \times \{0,1\} \times \mathbb{R})^n$. Let $\mathcal{F}$ denote the product σ-algebra generated from $\{\mathcal{P}(\{1,2\}) \times \mathcal{P}(\{0,1\}) \times \mathcal{B}\}^n$, where $\mathcal{B}$ is the Borel σ-algebra on $\mathbb{R}$, and for any set A, $\mathcal{P}(A)$ denotes the power set of A. Let $\Omega' = \mathcal{P}(\{H_{01}, H_{02}, H_{0*}\})$, and let $\mathcal{F}' = \mathcal{P}(\Omega')$. We assume each multiple testing procedure M is a measurable map from $(\Omega, \mathcal{F})$ to $(\Omega', \mathcal{F}')$. In what follows, when a property of a multiple testing procedure holds except on an event $E \in \mathcal{F}$ with $P(E) = 0$, we say it occurs with probability 1.

Consider the multiple testing procedure M^{STD}, defined to be the standard, one-sided z-test at level α for H_{0*}, i.e., the test that pools all participants and rejects if the standardized difference between sample means in the treatment arm and control arm exceeds $\Phi^{-1}(1-\alpha)$. It is uniformly most powerful for H_{0*} when outcomes under treatment and control are normally distributed with known variances; this follows directly from Proposition 15.2 of (van der Vaart, 1998).

We say a multiple testing procedure M' dominates a procedure M if for any $H \subseteq \{H_{01}, H_{02}, H_{0*}\}$, M' rejects H (and possibly additional null hypotheses) whenever M rejects H, with probability 1.

3.4. *Definition of strong control of asymptotic, familywise Type I error rate.* We require that all our testing procedures strongly control the familywise Type I error rate, also called the studywide Type I error rate, as defined by Hochberg and Tamhane (1987). Regulatory agencies such as the U.S. Food and Drug Administration and the European Medicines Agency generally require studywide Type I error control for confirmatory randomized trials involving multiple hypotheses (FDA and EMEA, 1998).

For a given multiple testing procedure, class of distributions $\mathcal{Q}'$, and sample size n, define the worst-case, familywise Type I error rate to be

$$\sup_{Q \in \mathcal{Q}'} P_{Q,n}\left(\text{at least one true null hypothesis is rejected}\right), \tag{3.6}$$

where $P_{Q,n}$ is the probability distribution resulting from outcome data being generated according to Q, at sample size n. We say a multiple testing procedure strongly controls the familywise Type I error rate at level α over $\mathcal{Q}'$ if for all sample sizes

n, (3.6) is at most α. We say a multiple testing procedure strongly controls the *asymptotic*, familywise Type I error rate at level α, over $\mathcal{Q}'$, if

$$\limsup_{n\to\infty} \sup_{Q\in\mathcal{Q}'} P_{Q,n}\left(\text{at least one true null hypothesis is rejected}\right) \leq \alpha. \tag{3.7}$$

For concreteness, we focus in what follows on the commonly used significance level $\alpha = 0.05$.

Strong control of the asymptotic, familywise Type I error rate as defined in (3.7) is desirable since it implies for any $\epsilon > 0$, there is a sample size N_ϵ that suffices to guarantee the familywise Type I error rate is at most $0.05 + \epsilon$, no matter what the data generating distribution (among those in $\mathcal{Q}'$). In contrast, under the following weaker, pointwise condition:

$$\sup_{Q\in\mathcal{Q}'} \limsup_{n\to\infty} P_{Q,n}\left(\text{at least one true null hypothesis is rejected}\right) \leq \alpha, \tag{3.8}$$

such a sample size may not exist. Though we focus on strong control of the asymptotic, familywise Type I error rate in the uniform sense as defined in (3.7), we also show in Section C.1 of the Supplementary Material that our main result, Theorem 4.1, still holds if we replace this by the pointwise condition (3.8).

3.5. *Subpopulation-specific and overall population z-statistics.* For subpopulation one, subpopulation two, and the overall population, respectively, define the following z-statistics:

$$Z_1 = \frac{\sum_{i:S_i=1}\{Y_iA_i - Y_i(1-A_i)\}}{\sigma_1(Q)(np_1)^{1/2}}, Z_2 = \frac{\sum_{i:S_i=2}\{Y_iA_i - Y_i(1-A_i)\}}{\sigma_2(Q)(np_2)^{1/2}}, \tag{3.9}$$

$$Z_* = \frac{1}{\sigma_*(Q)n^{1/2}}\sum_{i=1}^{n}\{Y_iA_i - Y_i(1-A_i)\}, \tag{3.10}$$

where for each $s \in \{1,2\}$, $\sigma_s^2(Q) = \left\{\sigma^2(Q_{s0}) + \sigma^2(Q_{s1})\right\}/2$, and $\sigma_*^2(Q) = p_1\sigma_1^2(Q) + p_2\sigma_2^2(Q)$.

For each $j \in \{*,1,2\}$, it follows that Z_j has variance 1. Also, for each subpopulation $s \in \{1,2\}$, it follows that the correlation between Z_s and Z_*, which we denote by ρ_s, satisfies

$$\rho_s = \left[p_s\sigma_s^2(Q)/\left\{p_1\sigma_1^2(Q) + p_2\sigma_2^2(Q)\right\}\right]^{1/2} > 0, \tag{3.11}$$

and that we have the following relationships:

$$\rho_1^2 + \rho_2^2 = 1 \text{ and } Z_* = \rho_1 Z_1 + \rho_2 Z_2. \tag{3.12}$$

For clarity of presentation, we assume throughout that the variances $\sigma^2(Q_{sa})$ are known, i.e., they are an input to any multiple testing procedure. However, we prove asymptotic versions of certain results (indicated in Sections 4.1 and 4.2) hold when the variances $\sigma^2(Q_{sa})$ are estimated by sample variances rather than assumed known, under the following additional condition: for some fixed $C' > 0$, for each $s \in \{1, 2\}, a \in \{0, 1\}$,

$$E_{Q_{sa}} \{Y - \mu(Q_{sa})\}^4 / \{\sigma^2(Q_{sa})\}^2 < C'. \tag{3.13}$$

This condition guarantees uniform convergence of sample variances to the population variances $\sigma^2(Q_{sa})$.

3.6. *Optimality criteria.* Let $\mathcal{Q}_N$ denote the class of data generating distributions $Q \in \mathcal{Q}$ in which each outcome distribution Q_{sa} is normally distributed. Let $\mathcal{C}$ denote the class of all multiple testing procedures for $\{H_{01}, H_{02}, H_{0*}\}$ that strongly control the familywise Type I error rate at level 0.05 over $\mathcal{Q}_N$. This includes but is not limited to procedures based on the closure principle of Marcus, Peritz and Gabriel (1976), or procedures based on partitioning as in Finner and Strassburger (2002), for example.

We say a multiple testing procedure $M \in \mathcal{C}$ is uniformly most powerful for simultaneously rejecting H_{0*} and at least one subpopulation null hypothesis, if for all $Q \in \mathcal{Q}_N$ for which H_{0*} is false and all $n > 0$, it satisfies

$$\begin{aligned} & P_{Q,n}(M \text{ rejects } H_{0*} \text{ and at least one of } H_{01}, H_{02}) \\ = & \sup_{M' \in \mathcal{C}} P_{Q,n}(M' \text{ rejects } H_{0*} \text{ and at least one of } H_{01}, H_{02}). \end{aligned} \tag{3.14}$$

For conciseness, we say a multiple testing procedure $M \in \mathcal{C}$ is uniformly most powerful for (3.14) to mean for all $Q \in \mathcal{Q}_N$ for which H_{0*} is false and all $n > 0$, it achieves the supremum (3.14). We give two different procedures that are uniformly most powerful for (3.14), in Section 4.1. This shows there is not a unique procedure that is uniformly most powerful for (3.14).

Consider the following properties:

A. Whenever the null hypothesis H_{0*} for the overall population is rejected, at least one subpopulation null hypothesis is rejected, with probability 1.
B. The probability of rejecting the null hypothesis H_{0*} is at least that of M^{STD}, i.e., the standard, one-sided z-test of H_{0*}, at level 0.05, for any data generating distribution.
C. Strong control of the familywise Type I error rate at level 0.05 over $\mathcal{Q}_N$.

It follows from Theorem 3.2.1 of Lehmann and Romano (2005) that having properties B and C is equivalent to the following: rejecting H_{0*} if and only if $Z_* > \Phi^{-1}(0.95)$, with probability 1.

In Section C.4 of the Supplementary Material, we prove for the case of two subpopulations:

THEOREM 3.1. *Any multiple testing procedure in $\mathcal{C}$ is uniformly most powerful for (3.14) if and only if it has properties A and B.*

4. Uniformly most powerful tests. We present results for tests of H_{01}, H_{02}, and H_{0*} in Section 4.1. In Section 4.2, we give results for the null hypotheses of zero mean treatment effect, and corresponding two-sided alternative hypotheses.

4.1. *One-sided hypotheses and tests.* Let M^{UMP} denote the following multiple testing procedure:

> Define S to be subpopulation one if $Z_1 - (3/4)\rho_1 \geq Z_2 - (3/4)\rho_2$, and subpopulation two otherwise. If $Z_* > \Phi^{-1}(0.95)$, reject H_{0*} and H_{0S}.

In Section C of the Supplementary Material, we prove:

THEOREM 4.1. *M^{UMP} satisfies the following:*

i. Properties A, B, and C from Section 3.6. Furthermore, it strongly controls the asymptotic, familywise Type I error rate at level 0.05 over $\mathcal{Q}$. This also holds if we replace $Z_*, Z_1, Z_2, \rho_1, \rho_2$ by corresponding quantities in which the variances $\sigma^2(Q_{sa})$ are estimated by sample variances rather than assumed known, under the additional condition (3.13).

ii. It is uniformly most powerful for (3.14).

Properties A and B follow directly from the definition of M^{UMP}. The main difficulty in proving Theorem 4.1 is showing M^{UMP} has property C. In Section 8, we explain why this is a challenging problem, and present our method for solving it and proving the other claims in part (i). We prove part (ii) follows from part (i) and Theorem 3.1, in Section C.5 of the Supplementary Material.

In the special case that $\rho_1 = \rho_2$, the procedure M^{UMP} reduces to the simpler procedure that, when $Z_* > \Phi^{-1}(0.95)$, rejects the overall population null hypothesis and the null hypothesis for the subpopulation with larger z-statistic. We denote this simpler procedure by M_0. This special case occurs, for example, when each subpopulation makes up exactly half the overall population and the variances $\sigma^2(Q_{sa})$ are all equal. However, when the subpopulations have different sizes or when these variances differ, M_0 can fail to strongly control the familywise Type I error rate, unlike the procedure M^{UMP}.

We now describe the intuition for how M^{UMP} reduces the worst-case, familywise Type I error rate, compared to the simpler procedure M_0. For clarity, we restrict to data generating distributions where outcomes are normally distributed.

We focus on the scenarios where the familywise Type I error rate of M_0 can exceed 0.05. This cannot happen if H_{0*} is true, since M_0 makes a Type I error only if $Z_* > \Phi^{-1}(0.95)$, and under H_{0*} this happens with probability at most 0.05. A familywise Type I error cannot occur if both H_{01}, H_{02} are false, since then H_{0*} is false as well, making a Type I error impossible.

The remaining case is the class of data generating distributions, denoted by $\tilde{\mathcal{Q}}$, for which H_{0*} is false and a single subpopulation null hypothesis, call it H_{01} without loss of generality, is false as well. Then M_0 only makes a Type I error when it rejects H_{02}, which occurs if both $Z_* > \Phi^{-1}(0.95)$ and $Z_2 - Z_1 > 0$. Direct computation shows this occurs with probability exceeding 0.05 only when the correlation between Z_* and $Z_2 - Z_1$, which equals $\rho_2 - \rho_1$, is positive. The procedure M^{UMP} raises the threshold for rejecting H_{02} in such cases, and therefore has a lower Type I error probability than M_0. The tradeoff is M^{UMP} has a higher Type I error probability than M_0 for $Q \in \tilde{\mathcal{Q}}$ when $\rho_2 - \rho_1 < 0$. However, since the Type I error probability for M_0 exceeds 0.05 only in the former case, this tradeoff reduces the *worst-case* Type I error probability over all $Q \in \tilde{\mathcal{Q}}$. We further explain this tradeoff, and how we selected the constant $3/4$ in the procedure M^{UMP}, in Section C.6 of the Supplementary Material.

We now show how to augment the procedure M^{UMP} to allow simultaneous rejection of all three null hypotheses H_{0*}, H_{01}, H_{02} in some cases, while still having all of the properties in Theorem 4.1. We consider each possible value of $\rho_1 \in (0, 1)$ separately, and use a threshold function $a(\rho_1)$ that we describe below. The augmented procedure $M^{\text{UMP+}}$, where the new part is in italics, is:

> Define S to be subpopulation one if $Z_1 - (3/4)\rho_1 \geq Z_2 - (3/4)\rho_2$, and subpopulation two otherwise. *If Z_1 and Z_2 are both greater than $a(\rho_1)$, reject all three null hypotheses H_{0*}, H_{01}, H_{02}.* Else, if $Z_* > \Phi^{-1}(0.95)$, reject H_{0*} and H_{0S}.

For each value $\rho_1' \in (0, 1)$, we set the threshold value $a(\rho_1')$ to be the smallest such that we can prove $M^{\text{UMP+}}$ strongly controls the asymptotic, familywise Type I error rate over $\{Q \in \mathcal{Q} : \rho_1(Q) = \rho_1'\}$ at level 0.05. In Section D of the Supplementary Material, we give an algorithm to compute $a(\rho_1)$, and plot the function $a(\rho_1)$. The set of values $\{a(\rho_1) : \rho_1 \in (0, 1)\}$ ranges between 1.92 and 2.19, with the minimum occurring at $\rho_1 = 2^{-1/2}$, i.e., where $\rho_1 = \rho_2$.

All of the above results hold for any value of $p_1 : 0 < p_1 < 1$; that is, regardless of the fraction p_1 of the overall population in subpopulation one, our procedure M^{UMP} has properties A, B, and C, and is uniformly most powerful for (3.14). Since $M^{\text{UMP+}}$ dominates M^{UMP}, we have $M^{\text{UMP+}}$ is also uniformly most powerful for (3.14).

4.2. *Two-sided tests.* We consider testing the null hypotheses of zero mean treatment effect. That is, we consider the null hypotheses $H_{0*}^T, H_{01}^T, H_{02}^T$ defined by replacing each occurrence of "≤ 0" by "$= 0$" in (3.2), (3.3), and (3.4). The corresponding alternative hypotheses are the complements of each of these null hypotheses. We define the multiple testing procedure M^{TS} (where "TS" abbreviates "two-sided") as follows:

> Define S' to be subpopulation one if $|Z_1| - (1/2)\rho_1 \geq |Z_2| - (1/2)\rho_2$, and subpopulation two otherwise. If $|Z_*| > \Phi^{-1}(0.975)$, reject H_{0*}^T and $H_{0S'}^T$.

This procedure has properties analogous to A, B, and C, which we describe next. Define the standard, two-sided z-test for H_{0*}^T at level α to be the test that pools all subjects and rejects for large absolute values of the standardized difference between sample means in the treatment and control arms, i.e., when $|Z_*| > \Phi^{-1}(1-\alpha/2)$. Define properties A^T, B^T, and C^T, as follows:

- A^T. Whenever the null hypothesis H_{0*}^T is rejected, at least one of the subpopulation null hypotheses H_{01}^T, H_{02}^T is rejected, with probability 1.
- B^T. The probability of rejecting H_{0*}^T is at least that of the standard, two-sided z-test of H_{0*}^T at level 0.05, for any data generating distribution.
- C^T. Strong control of the asymptotic, familywise Type I error rate at level 0.05 over $\mathcal{Q}_N$.

It follows from the definition of M^{TS} that it has properties A^T and B^T. We prove in Section E of the Supplementary Material that it has property C^T.

Just as there is no uniformly most powerful test for H_{0*}^T (as shown, e.g., in (van der Vaart, 1998, Section 15)), there is no uniformly most powerful test for simultaneously rejecting H_{0*}^T and at least one of H_{01}^T, H_{02}^T. We prove this in Section E of the Supplementary Material. However, the procedure M^{TS} has a maximin optimality property that we define next.

Let $\Delta_{\min} > 0$ denote the magnitude of the minimum, clinically meaningful difference between means under treatment versus control. Let v represent an upper bound on the variances $\sigma^2(Q_{sa})$. Define the class of alternatives ω to be those $Q \in \mathcal{Q}_N$ satisfying both:

i. The mean treatment effect $\mu(Q_{s1}) - \mu(Q_{s0}) \geq \Delta_{\min}$ for both subpopulations $s \in \{1,2\}$, or the treatment effect $\mu(Q_{s1}) - \mu(Q_{s0}) \leq -\Delta_{\min}$ for both subpopulations $s \in \{1,2\}$.
ii. For each study arm $a \in \{0,1\}$, the variance of the outcome is the same for both subpopulations and is at most v, i.e., $\sigma^2(Q_{1a}) = \sigma^2(Q_{2a}) \leq v$.

We say a multiple testing procedure is maximin optimal over ω if for each sample size n, it maximizes

$$\inf_{Q \in \omega} P_{Q,n}(\text{reject } H_{0*}^T \text{ and at least one of } H_{01}^T, H_{02}^T), \tag{4.1}$$

among all multiple testing procedures for $H_{0*}^T, H_{01}^T, H_{02}^T$ with property C^T. We prove the following in Section E of the Supplementary Material:

THEOREM 4.2. *M^{TS} has the following properties:*

i. Properties A^T, B^T, and C^T. Furthermore, it strongly controls the asymptotic, familywise Type I error rate at level 0.05 over $\mathcal{Q}$. This also holds if we replace $Z_*, Z_1, Z_2, \rho_1, \rho_2$ by corresponding quantities in which the variances $\sigma^2(Q_{sa})$ are estimated by sample variances rather than assumed known, under the additional condition (3.13).

ii. It maximizes (4.1) among all multiple testing procedures with property C^T.

The proof of the above theorem uses our general method described in Section 8 in combination with a method of Romano, Shaikh and Wolf (2011) for proving maximin optimality of consonant procedures.

5. Relationship between property A and consonance. We first define a property related to consonance. Coherence is the property for multiple testing procedures that whenever an intersection of null hypotheses is not rejected, neither is the intersection of any subset of these null hypotheses. Sonnemann and Finner (1988) show that any multiple testing procedure can be converted into a coherent procedure that rejects the same null hypotheses (and possibly more) without affecting the familywise Type I error rate. In our context, by definition, any coherent procedure that rejects H_{0*} must reject $H_{0*} \cap H_{01} \cap H_{02}$. By (3.5), this intersection equals $H_{01} \cap H_{02}$. Thus, any coherent procedure that rejects H_{0*} must reject $H_{01} \cap H_{02}$.

There are multiple variants of the definition of consonance, as described by Brannath and Bretz (2010). Here, we call a testing procedure consonant if for each non-empty subset $J \subseteq \{*, 1, 2\}$, rejection of the intersection null hypothesis $\cap_{j \in J} H_{0j}$ implies rejection of at least one elementary null hypothesis $H_{0j'}$ for $j' \in J$. In particular, a consonant procedure that rejects $H_{01} \cap H_{02}$ must reject at least one of the elementary null hypotheses H_{01}, H_{02}. Combining this with the claim at the end of the previous paragraph, we have that any consonant, coherent procedure must reject at least one of H_{01}, H_{02} whenever it rejects H_{0*}. That is, any consonant, coherent procedure must have property A from Section 3.6. The converse holds as well, when restricting to coherent multiple testing procedures with properties B and C, and focusing on the class of normally distributed data generating distributions $\mathcal{Q}_N$ defined above; this is expressed in the following theorem, proved in Section F of the Supplementary Material:

THEOREM 5.1. *Consider any coherent multiple testing procedure M for the family of null hypotheses $\mathcal{H}$. Assume M satisfies properties B and C for the class of distributions $\mathcal{Q}_N$. Then for any $Q \in \mathcal{Q}_N$, property A holds for M with probability 1 (under Q) if and only if M is consonant with probability 1 (under Q).*

As described in Section F of the Supplementary Material, this close link between property A and consonance arises from the relationship (3.5) between subpopulation and overall population null hypotheses.

Romano, Shaikh and Wolf (2011) give a general algorithm for constructing a consonant procedure from a multiple testing procedure that is not consonant. However, the resulting procedure will not necessarily satisfy property A. This is due to a subtle difference in the definition of consonance here and in (Romano, Shaikh and Wolf, 2011), which we describe in Section F of the Supplementary Material.

6. Existing multiple testing procedures. The fixed sequence method of Maurer, Hothorn and Lehmacher (1995) can be applied to the ordering (H_{0*}, H_{01}, H_{02}). The resulting procedure, denoted M^{FS}, involves proceeding along this ordering, testing each H_{0j} using the test $Z_j > \Phi^{-1}(0.95)$, until the first failure to reject, at which point the procedure stops. This procedure may be desirable when one has prior evidence that the treatment effect is likely to be stronger in subpopulation one compared to subpopulation two; however, a downside is that H_{01} must be rejected before H_{02} can even be considered. This downside is especially relevant when it is not known with certainty which subpopulation will benefit more from an experimental treatment, or when the subpopulation proportions and variances make a treatment benefit more difficult to detect in subpopulation one compared to subpopulation two (e.g., when $p_1 < p_2$ or $\sigma_1(Q) > \sigma_2(Q)$).

The following multiple testing procedure, denoted M^{R}, was given in the case of $p_1 = 1/2$ by (Rosenbaum, 2008, Section 2):

> M^{R}: If $Z_* > \Phi^{-1}(0.95)$, reject H_{0*} as well as each subpopulation null hypothesis H_{0s}, $s \in \{1, 2\}$, for which $Z_s > \Phi^{-1}(0.95)$.

Rosenbaum (2008) shows this procedure strongly controls the familywise Type I error rate at level 0.05 over $\mathcal{Q}_N$. A straightforward extension of that proof shows the result holds for any $p_1 \in (0, 1)$. By construction, M^{R} has property B. However, it does not have property A. That is, the procedure may reject the overall population null hypothesis without rejecting any subpopulation null hypothesis. This follows since $Z_* > \Phi^{-1}(0.95)$ does not imply at least one of Z_1, Z_2 is greater than $\Phi^{-1}(0.95)$. We give a data example in which $Z_* > \Phi^{-1}(0.95)$, but neither Z_1 nor Z_2 exceeds $\Phi^{-1}(0.95)$, in Section A of the Supplementary Material. Since M^{R} dominates M^{FS}, we only consider the former in the power comparison below.

Bergmann and Hommel (1988) give an improvement to the Holm step-down procedure for hypotheses that are logically related. As described by Hommel and Bernhard (1999), the procedure of Bergmann and Hommel (1988), here denoted M^{BH}, involves first specifying which subsets of elementary null hypotheses are "exhaustive." For any index set $J \subseteq \{*, 1, 2\}$, the subset $\{H_{0j}, j \in J\}$ is defined to be exhaustive if there exists a data generating distribution under which all and only the null hypotheses in this subset are true. In our problem, all subsets are exhaustive except $\{H_{01}, H_{02}\}$ and the singleton $\{H_{0*}\}$, since the relationship (3.5) implies that whenever H_{01}, H_{02} are both true also H_{0*} is true, and whenever H_{0*} is true at least one of H_{01}, H_{02} is true. The procedure M^{BH} rejects the null hypotheses with indices $\{*, 1, 2\} \setminus A$, where A is defined as the union of all subsets $J \subseteq \{*, 1, 2\}$ that satisfy:

$$\{H_{0j}, j \in J\} \text{ is exhaustive and } \max\{Z_j : j \in J\} < \Phi^{-1}(1 - 0.05/|J|).$$

We show in Section B of the Supplementary Material that this procedure has neither properties A nor B.

Song and Chi (2007) and Alosh and Huque (2009) designed multiple testing procedures involving the overall population and a single, prespecified subpopulation s^*. Here, in contrast, we are interested in the larger family of hypotheses including the subpopulation complementary to s^*. To tailor the procedure of Song and Chi (2007) to our context, we augment it to additionally allow rejection for the complementary subpopulation, without any loss in power for H_{0*} or for H_{0s^*}, and while maintaining strong control of the familywise Type I error rate. We denote the augmented procedure by $M^{\mathrm{SC}+,s^*}$, which, for prespecified thresholds $\alpha_0, \alpha_1, \alpha_2$ satisfying $0 \le \alpha_0 < 0.05 < \alpha_1 \le 1$, and $0 \le \alpha_2 \le 1$, is defined as follows:

> If $Z_* > \Phi^{-1}(1 - \alpha_0)$, reject H_{0*} as well as each subpopulation null hypothesis H_{0s}, $s \in \{1, 2\}$, for which $Z_s > \Phi^{-1}(1 - 0.05)$.
> If $\Phi^{-1}(1 - \alpha_0) \ge Z_* > \Phi^{-1}(1 - \alpha_1)$ and $Z_{s^*} > \Phi^{-1}(1 - \alpha_2)$, then reject H_{0s^*}, and if in addition $Z_* > \Phi^{-1}(1 - 0.05)$ then reject H_{0*}.

The original procedure of Song and Chi (2007), which we denote by M^{SC,s^*}, is the same as above except it does not allow rejection of the null hypothesis complementary to s^*. We chose $\alpha_0 = 0.045$ and $\alpha_1 = 0.1$. We then used the method of Song and Chi (2007) to compute, for each scenario we consider, the largest α_2 (which depends on p_1) such that the above procedure strongly controls the familywise Type I error rate at level 0.05. For $p_1 = 1/2$, we have $\alpha_2 = 0.023$; for $p_1 = 3/4$, we have $\alpha_2 = 0.025$. We show in Section B of the Supplementary Material that $M^{\mathrm{SC}+,s^*}$ strongly controls the asymptotic, familywise Type I error rate at level 0.05, and in general it has neither properties A nor B. The procedure M^{SC,s^*} of Song and Chi (2007) has similar performance to a procedure of Alosh and Huque (2009), so we only include the former in our comparison below.

7. Simulations to assess power and Type I error.

7.1. *Power comparison.* We compare the power of $M^{\text{UMP+}}$, M^{R}, M^{BH}, $M^{\text{SC},1}$, $M^{\text{SC+},1}$, and $M^{\text{SC+},2}$. We consider a wide variety of data generating distributions Q, and give full results in Section B.2 of the Supplementary Material. Here, we focus on five representative cases. In each case, we set each outcome distribution Q_{sa} to be normally distributed with all variances $\sigma^2(Q_{sa})$ equal to a common value denoted by σ^2.

We consider two types of scenarios. In the first, we set the mean treatment effect $\mu(Q_{s1}) - \mu(Q_{s0})$ for each subpopulation s to be the same value $\Delta\mu > 0$, where $\Delta\mu$ (defined in Section B.1 of the Supplementary Material) is the value at which the standard, one-sided z-test of H_{0*} has 80% power. In the second type of scenario, we set only subpopulation one to benefit from treatment, by letting $\mu(Q_{11}) - \mu(Q_{10}) = \Delta\mu$ and $\mu(Q_{21}) - \mu(Q_{20}) = 0$. In each scenario, we consider several values of p_1.

We say a multiple testing procedure rejects at least the set of null hypotheses $\mathcal{G}$, if it rejects all of these null hypotheses and possibly more. For each testing procedure and data generating distribution, we ran 10^6 Monte Carlo iterations and recorded the empirical probabilities of rejecting each subset of null hypotheses. These are given in Table 1, rounded to the nearest percent. In scenario 1, all null hypotheses are false; in scenario 2, only H_{0*} and H_{01} are false, and so power is given only for subsets involving these null hypotheses.

In all cases, $M^{\text{UMP+}}$ has the maximum power for rejecting H_{0*} and at least one false subpopulation null hypothesis. It is precisely this goal that $M^{\text{UMP+}}$ is designed for. Also, in all cases, $M^{\text{UMP+}}$ has the maximum power for rejecting at least the overall population null hypothesis H_{0*}.

The augmented version $M^{\text{SC+},1}$ of the procedure $M^{\text{SC},1}$ of Song and Chi (2007) has substantially more power (up to 52% more) than $M^{\text{SC},1}$ to reject H_{02} in scenario 1. This is not surprising, since $M^{\text{SC},1}$ was designed for testing only $\{H_{0*}, H_{01}\}$, rather than $\{H_{0*}, H_{01}, H_{02}\}$ as considered here.

The procedure M^{R} has the same power as $M^{\text{UMP+}}$ to reject H_{0*}, and has 5-6% less power to simultaneously reject H_{0*} and at least one of H_{01}, H_{02} in scenario 1. However, M^{R} is roughly equal to or improves on the power of $M^{\text{UMP+}}$ in all other cases, with a large improvement (up to 14%) in scenario 1 in the power to reject all three null hypotheses.

It would be ideal to construct a multiple testing procedure in our setting having both the advantages of M^{R} (including dominating M^{FS}, and having properties B and C) and the advantages of $M^{\text{UMP+}}$ (including properties A, B, C, and being uniformly most powerful for (3.14)). Unfortunately, this is not possible, in that:

THEOREM 7.1. *No multiple testing procedure for $\{H_{01}, H_{02}, H_{0*}\}$ simultaneously has properties A, B, and C, and dominates the fixed sequence procedure*

TABLE 1

Power Comparison. Each cell reports the probability (as a percent) that the procedure in that row rejects at least the set of null hypotheses corresponding to that column. The column heading "H_{0*}+sub" means H_{0*} and at least one of H_{01}, H_{02}; "all" means all three null hypotheses.

Scenario 1: Both subpopulations benefit equally from treatment

	$p_1 = 1/2$					$p_1 = 3/4$				
	H_{0*}	H_{0*} +sub	H_{0*} $+H_{01}$	H_{0*} $+H_{02}$	all	H_{0*}	H_{0*} +sub	H_{0*} $+H_{01}$	H_{0*} $+H_{02}$	all
$M^{\text{UMP+}}$	80	80	49	49	19	80	80	62	28	10
M^{R}	80	74	52	52	30	80	75	67	32	24
M^{BH}	66	65	48	48	30	66	66	60	29	24
$M^{\text{SC},1}$	79	52	52	0	0	79	67	67	0	0
$M^{\text{SC}+,1}$	79	74	52	52	30	79	75	67	32	24
$M^{\text{SC}+,2}$	79	74	52	52	30	79	75	66	32	24

Scenario 2: Only subpopulation one benefits from treatment

	$p_1 = 1/2$			$p_1 = 2/3$			$p_1 = 3/4$		
	H_{0*}	H_{0*} $+H_{01}$	H_{01}	H_{0*}	H_{0*} $+H_{01}$	H_{01}	H_{0*}	H_{0*} $+H_{01}$	H_{01}
$M^{\text{UMP+}}$	34	31	31	51	47	47	59	55	55
M^{R}	34	30	30	51	47	47	59	55	55
M^{BH}	22	20	38	36	35	49	44	43	54
$M^{\text{SC},1}$	34	29	36	50	46	52	58	55	60
$M^{\text{SC}+,1}$	34	29	36	50	46	52	58	55	60
$M^{\text{SC}+,2}$	33	28	28	48	45	45	57	54	54

M^{FS}. Also, no multiple testing procedure for $\{H_{01}, H_{02}, H_{0*}\}$ with property C simultaneously dominates M^{FS} and is uniformly most powerful for (3.14).

We prove the above theorem in Section G of the Supplementary Material. The upshot is that there is a tradeoff between dominating M^{FS} on the one hand, and being uniformly most powerful for rejecting H_{0*} and at least one subpopulation null hypothesis on the other hand, for procedures with property C.

The above scenarios capture the main features of the extensive simulations depicted in Figures 2 and 3 of Section B.2 of the Supplementary Material. The one exception is the scenario, not represented above, but depicted in Figure 2, where the mean treatment effect is negative for one subpopulation and positive for the other. This is called a qualitative interaction, as opposed to a quantitative interaction. Since these treatment effects partially cancel out for the overall population, all the procedures above have relatively low power for rejecting H_{0*}. In such scenarios, M^{BH} has the most power to reject at least one of H_{01}, H_{02}, but this power is not very large. All the above procedures, including our new procedures and the existing procedures we compared against, are for situations where it is suspected that there may be quantitative, rather than qualitative, interactions. In cases where one suspects a qualitative interaction, other methods should be considered.

7.2. *Familywise Type I error rate.* In the special case where each outcome distribution Q_{sa} is normally distributed, the familywise Type I error rate of $M^{\text{UMP+}}$ is at most 0.05, at any sample size. In general, the familywise Type I error guarantee in Theorem 4.1 is asymptotic, as sample size goes to infinity, as defined in Section 3.4. We did extensive simulations based on skewed and heavy-tailed distributions in $\tilde{\mathcal{Q}}$, with sample sizes from 50 to 500 participants. These are described in Section B.3 of the Supplementary Material. As a benchmark for how challenging each data generating distribution $Q \in \tilde{\mathcal{Q}}$ is, we computed the Type I error of the standard, one-sided z-test for H_{0*} under $c(Q)$, where $c(Q)$ is the distribution resulting from centering each component distribution of Q to have mean 0. For each data generating distribution $Q \in \tilde{\mathcal{Q}}$ we simulated from, the familywise Type I error rate of $M^{\text{UMP+}}$ under Q was never more than the Type I error of the standard, one-sided z-test under $c(Q)$.

8. General method for finding the least favorable distribution. The main challenge in showing the procedures M^{UMP} and $M^{\text{UMP+}}$ strongly control the familywise Type I error rate at level 0.05 is to identify, for each procedure, the least favorable data generating distribution, that is, the distribution that maximizes the familywise Type I error rate. To show why this is not trivial, consider the global null distribution, defined as the $Q \in \mathcal{Q}_N$ for which each subpopulation's mean treatment effect $\mu(Q_{s1}) - \mu(Q_{s0})$ is zero, and all variances equal 1. It is not clear,

a priori, whether this distribution is least favorable for M^{UMP}. In fact, we show in Section C.6 of the Supplementary Material that for the simpler procedure M_0 defined in Section 4.1, the least favorable distribution is not the global null distribution, and the resulting familywise Type I error rate exceeds 0.05. The proof of Theorem 4.1 involves showing this does not occur for M^{UMP}. We present the main steps of the proof below. The full proof is given in Section C of the Supplementary Material.

Recall that we place only the minimal constraints from Section 3.2 on the means and variances of the outcome distribution for each subpopulation and treatment arm, and put no constraints on the proportions of the overall population in each subpopulation. Finding the least favorable distribution involves optimizing over all possible values of these to find the distribution that maximizes the familywise Type I error rate. This is a nonlinear optimization problem that is either difficult or impossible to analytically solve. However, we can solve it to any desired precision with a combination of analytical arguments and intensive computation. Though this approach, broadly speaking, is similar to that in (Rosenblum and van der Laan, 2011), the specifics are quite different. Each new multiple testing procedure requires new analytic arguments for transforming the corresponding, difficult optimization problem into a small set of computationally tractable problems. We explain the simplest case below.

To compute the least favorable distribution for procedure M^{UMP}, we first compute this within the subclass of data generating distributions $Q \in \mathcal{Q}$ for which H_{0*} is true, which we denote by $\mathcal{Q}_*$. This is the class of $Q \in \mathcal{Q}$ for which, for all n, we have $E_{Q,n}(Z_*) \leq 0$. For $Q \in \mathcal{Q}$ for which H_{0*} is true, a familywise Type I error occurs under M^{UMP} if and only if $Z_* > \Phi^{-1}(0.95)$. By the uniform central limit theorem of Götze (1991) and the assumptions in Section 3.2, we have

$$\lim_{n\to\infty} \sup_{Q\in\mathcal{Q}_*, t\in\mathbb{R}} |P_{Q,n}\{Z_* - E_{Q,n}(Z_*) > t\} - \Phi(-t)| = 0. \tag{8.1}$$

This implies, taking $t = \Phi^{-1}(0.95)$, that for the class $\mathcal{Q}_*$, the asymptotic, familywise Type I error rate is at most 0.05.

It remains to consider the class of data generating distributions $Q \in \mathcal{Q}$ for which H_{0*} is false. If H_{0*}, H_{01}, and H_{02} are all false, the familywise Type I error rate equals 0. It therefore suffices to consider $Q \in \mathcal{Q}$ for which H_{0*} is false, and exactly one subpopulation null hypothesis, say H_{01} without loss of generality, is false. Recall this class is denoted by $\tilde{\mathcal{Q}}$.

For each $Q \in \tilde{\mathcal{Q}}$ and $n > 0$, define the centered statistics $Z_j^c = Z_j - E_{Q,n}Z_j$, for each $j \in \{*, 1, 2\}$. Define $\rho'(Q) = 2^{-1/2}\{\rho_2(Q) - \rho_1(Q)\}$. For any $Q \in \tilde{\mathcal{Q}}$

and $n > 0$, the probability M^{UMP} makes a familywise Type I error is

$$
\begin{aligned}
(8.2) \qquad & P_{Q,n}\left\{Z_* > \Phi^{-1}(0.95), Z_2 - (3/4)\rho_2(Q) > Z_1 - (3/4)\rho_1(Q)\right\} \\
= \; & P_{Q,n}\left\{Z_*^c > \Phi^{-1}(0.95) - E_{Q,n}Z_*, \right. \\
& \qquad \left. (Z_2^c - Z_1^c)/\sqrt{2} > (3/4)\rho'(Q) - E_{Q,n}(Z_2 - Z_1)/\sqrt{2}\right\} \\
(8.3) \quad = \; & P_{Q,n}\left\{\left(Z_*^c, (Z_2^c - Z_1^c)/\sqrt{2}\right) \in (\lambda_1(Q,n), \infty) \times (\lambda_2(Q,n), \infty)\right\},
\end{aligned}
$$

where

$$
\begin{aligned}
\lambda_1(Q,n) &= \Phi^{-1}(0.95) - E_{Q,n}Z_* \\
(8.4) \qquad &= \Phi^{-1}(0.95) - \rho_1(Q)E_{Q,n}Z_1 - \rho_2(Q)E_{Q,n}Z_2; \\
(8.5) \qquad \lambda_2(Q,n) &= (3/4)\rho'(Q) - E_{Q,n}\left(Z_2 - Z_1\right)/\sqrt{2}.
\end{aligned}
$$

Let $G(y)$ denote a bivariate normal random vector with zero mean and covariance matrix with 1s on the main diagonal and y off the main diagonal. Applying the uniform central limit theorem of Götze (1991), and that the covariance of $(Z_*^c, (Z_2^c - Z_1^c)/\sqrt{2})$ is $\rho'(Q)$, we have

$$
(8.6) \qquad \lim_{n\to\infty} \sup_{Q\in\tilde{\mathcal{Q}}, A\in\mathcal{A}} \left| P_{Q,n}\left\{\left(Z_*^c, (Z_2^c - Z_1^c)/\sqrt{2}\right) \in A\right\} - P\left\{G(\rho'(Q)) \in A\right\}\right| = 0,
$$

where $\mathcal{A}$ denotes the set of all Borel measurable, convex subsets of $\mathbb{R}^2$. The above display, combined with the equality of (8.2) and (8.3), implies the following characterization of the asymptotic, worst-case over $Q \in \tilde{\mathcal{Q}}$, familywise Type I error:

$$
\begin{aligned}
& \limsup_{n\to\infty} \sup_{Q\in\tilde{\mathcal{Q}}} P_{Q,n}\left\{Z_* > \Phi^{-1}(0.95), Z_2 - \frac{3}{4}\rho_2(Q) > Z_1 - \frac{3}{4}\rho_1(Q)\right\} \\
(8.7) \quad & = \limsup_{n\to\infty} \sup_{Q\in\tilde{\mathcal{Q}}} P\left\{G(\rho'(Q)) \in (\lambda_1(Q,n), \infty) \times (\lambda_2(Q,n), \infty)\right\}.
\end{aligned}
$$

For any n, the term inside the $\limsup$ in (8.7) can be replaced by

$$
(8.8) \qquad \sup_{(x_1,x_2,x_3)\in A_n} P\left\{G(x_3) \in (x_1, \infty) \times (x_2, \infty)\right\},
$$

where A_n is the set of triples $\{(\lambda_1(Q,n), \lambda_2(Q,n), \rho'(Q))\}_{Q\in\tilde{\mathcal{Q}}}$. We have reduced the problem of computing the asymptotic, worst-case over $Q \in \tilde{\mathcal{Q}}$, familywise Type I error rate to the optimization problem (8.8).

In Section C.1 of the Supplementary Material, we precisely characterize the region A_n, and show A_n does not depend on n. We then give an algorithm and

R code to solve the non-convex optimization problem (8.8) to any desired accuracy. This is achieved by doing a grid search over A_n, where at each grid point $(x_1, x_2, x_3) \in A_n$ we compute the bivariate normal probability in (8.8) using the R package mvtnorm, and take the maximum found over all grid points. We then prove an analytic bound on the approximation error in the grid search, by bounding the gradient of $P\{G(x_3) \in (x_1, \infty) \times (x_2, \infty)\}$ with respect to (x_1, x_2, x_3) and using the mean value theorem. The result is that (8.8) is at most 0.0461, which by (8.7) implies the same for the asymptotic, familywise Type I error rate for M^{UMP} over $Q \in \tilde{\mathcal{Q}}$. Combining this with the above upper bound of 0.05 over the subclass $\mathcal{Q}_*$, we have the asymptotic, familywise Type I error rate for M^{UMP} over the class $\mathcal{Q}$ is at most 0.05.

Our method for solving the non-convex optimization problem (8.8) has the advantage that despite the presence of local optima, it is guaranteed to give a solution to any desired accuracy. This is an improvement over methods such as simulated annealing and the Nelder-Mead algorithm, which may get stuck in local optima. Our method relies on partitioning the overall optimization problem into computationally tractable smaller problems. It can be quite challenging to devise such a partition, and the bulk of Sections C and E of the Supplementary Material is devoted to constructing such partitions for our multiple testing procedures M^{UMP}, $M^{\text{UMP+}}$, and M^{TS}.

9. More than two subpopulations. In the case of two subpopulations, we exhibited procedures that are uniformly most powerful as in (3.14), and that have properties A, B, and C. This raises the hope this may be possible for $k > 2$ subpopulations. Consider the case where the overall population is partitioned into $k > 2$ subpopulations, in proportions $p_1, \ldots, p_k$. The definitions in Section 3 for two subpopulations naturally generalize to $k > 2$ subpopulations. Define the null hypotheses of no positive mean treatment effect in each subpopulation and in the overall population, respectively, as:

$$H_{0s} = \{Q \in \mathcal{Q} : \mu(Q_{s1}) - \mu(Q_{s0}) \leq 0\}, \text{ for each } s \in \{1, \ldots, k\};$$
$$H_{0*} = \left\{Q \in \mathcal{Q} : \sum_{s=1}^{k} p_s \{\mu(Q_{s1}) - \mu(Q_{s0})\} \leq 0\right\}.$$

In Section G of the Supplementary Material we prove:

THEOREM 9.1. *Consider any $k > 2$, and assume the overall population is partitioned into $k > 2$ subpopulations. No multiple testing procedure simultaneously has properties A, B, and C.*

To explain the key idea underlying the proof of Theorem 9.1, take the simplest case of $k = 3$ subpopulations and $p_1 = p_2 = p_3 = 1/3$. For each $j \in \{1, 2, 3\}$,

define $Q^{(j)}$ to be the data generating distribution in $\mathcal{Q}_N$ having mean treatment effect $\delta > 0$ for subpopulation j, and 0 for the remaining two subpopulations; we describe how δ is chosen below. Define $Q^{(0)}$ to have mean treatment effect 0 in all subpopulations. Set all variances $\sigma^2(Q_{sa}^{(j)}) = 1$. Consider any multiple testing procedure M with properties A and B. We prove in Section G of the Supplementary Material that the familywise Type I error rate of M exceeds 0.05 for at least one distribution $Q^{(j)}$ for $j \in \{0, 1, 2, 3\}$, under a certain choice of δ. We next explain the main steps in this proof.

Assume, for the sake of contradiction, that M has Type I error at most 0.05 for each $Q^{(j)}$. Since M has property B, with probability 1, it rejects H_{0*} whenever $Z_* > \Phi^{-1}(0.95)$. By property A, whenever this occurs, M must reject at least one subpopulation null hypothesis as well. For any $j \in \{1, 2, 3\}$, this leads to a Type I error under $Q^{(j)}$ unless the rejected null hypothesis is H_{0j}, i.e., the null hypothesis corresponding to the single subpopulation where there is a positive mean treatment effect δ. We prove in Section G of the Supplementary Material that for small values of $\delta > 0$, under $Q^{(j)}$, it is impossible to reliably pick out subpopulation j; we use this to show for a certain choice of δ, the familywise Type I error rate exceeds 0.05 for at least one of the $Q^{(j)}$. An extension of this idea is used to prove the general case of $k > 2$ subpopulations and any set of subpopulation proportions. We formalize the above argument in Section G of the Supplementary Material.

10. Discussion. Having constructed procedures that are uniformly most powerful for (3.14) allowed us determine what properties are possible (and impossible) to simultaneously achieve in our setting, and to demonstrate tradeoffs connected with these properties. We showed for two subpopulations that no procedure dominating M^{R} or M^{FS} can be uniformly most powerful for (3.14). Therefore, one has to choose which properties are most important when selecting a multiple testing procedure.

The power comparisons in Section 7 provide information that may be useful in selecting a multiple testing procedure. If it is strongly desired to guarantee rejecting at least one subpopulation null hypothesis whenever H_{0*} is rejected, $M^{\mathrm{UMP+}}$ could be useful. The procedure M^{R}, though lacking this property, has quite favorable overall performance in the power comparison in Section 7 and in the Supplementary Material; in particular, it has substantially greater power than $M^{\mathrm{UMP+}}$ for simultaneously rejecting all three null hypotheses. This can be an important consideration for use in practice. It is an open research problem to simultaneously optimize a weighted combination of (i) power for simultaneously rejecting all three null hypotheses, and (ii) power for rejecting H_{0*} and at least one subpopulation null hypothesis, each at certain alternatives of interest.

We compare our new multiple testing procedures to existing procedures in a data

example from a trial of trastuzumab to treat metastatic breast cancer in Section A of the Supplementary Material. In this example, our new procedures M^{UMP} and $M^{\text{UMP+}}$ reject the overall population null hypothesis and a subpopulation null hypothesis, while all the other procedures considered in this paper reject only the former or reject nothing.

Acknowledgments. I would like to acknowledge Tom Louis and Karen Bandeen-Roche for their helpful suggestions on this paper. This research and analysis was supported by contract number HHSF2232010000072C, entitled, "Partnership in Applied Comparative Effectiveness Science," sponsored by the Food and Drug Administration, Department of Health and Human Services. This publication's contents are solely the responsibility of the authors and do not necessarily represent the official views of the above agencies.

Supplementary Material. The Supplementary Material is available at:

`http://people.csail.mit.edu/mrosenblum/papers/ump.pdf`

In Section A of the Supplementary Material, we illustrate the methods compared above in a data example based on a randomized trial of treatments for metastatic breast cancer. Section B presents simulations comparing power and familywise Type I error rates. We prove Theorems 3.1 and 4.1 in Section C. Section D gives the algorithm for the threshold $a(\rho_1)$ used in the augmented procedure $M^{\text{UMP+}}$. We prove Theorem 4.2 for two-sided tests in Section E. We discuss the relationship between property A and consonance, and prove Theorem 5.1, in Section F. Proofs of Theorems 7.1 and 9.1, which show certain properties cannot be simultaneously satisfied by any multiple testing procedure, are given in Section G.

References.

ALOSH, M. and HUQUE, M. F. (2009). A flexible strategy for testing subgroups and overall population. *Statistics in Medicine* **28** 3–23.

BERGMANN, B. and HOMMEL, G. (1988). Improvements of general multiple test procedures for redundant systems of hypotheses. In *Multiple Hypothesenprüfung–Multiple Hypotheses Testing* (P. Bauer, G. Hommel and E. Sonnemann, eds.) 100-115. Springer, Berlin.

BITTMAN, R. M., ROMANO, J. P., VALLARINO, C. and WOLF, M. (2009). Optimal testing of multiple hypotheses with common effect direction. *Biometrika* **96** 399-410.

BRANNATH, W. and BRETZ, F. (2010). Shortcuts for locally consonant closed test procedures. *Journal of the American Statistical Association* **105** 660-669.

FDA and EMEA (1998). E9 statistical principles for clinical trials. *U.S. Food and Drug Administration: CDER/CBER. European Medicines Agency: CPMP/ICH/363/96. http://www.fda.gov/cder/guidance/index.htm*.

FINNER, H. and STRASSBURGER, K. (2002). The partitioning principle: a powerful tool in multiple decision theory. *Ann. Statist.* **Volume 30** 1194-1213.

GABRIEL, K. R. (1969). Simultaneous test procedures – some theory of multiple comparisons. *The Annals of Mathematical Statistics* **40** 224-250.

GÖTZE, F. (1991). On the rate of convergence in the multivariate CLT. *The Annals of Probability* **19** 724-739.

HOCHBERG, Y. and TAMHANE, A. C. (1987). *Multiple Comparison Procedures*. Wiley Interscience, New York.

HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6** 65–70.

HOMMEL, G. (1986). Multiple test procedures for arbitrary dependence structures. *Metrika* **33** 321-336.

HOMMEL, G. and BERNHARD, G. (1999). Bonferroni procedures for logically related hypotheses. *Journal of Statistical Planning and Inference* **82** 119 - 128.

KIRSCH, I., DEACON, B. J., HUEDO-MEDINA, T. B., SCOBORIA, A., MOORE, T. J. and JOHNSON, B. T. (2008). Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. *PLoS Med* **5** e45.

LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. ed. Springer, New York.

MARCUS, R., PERITZ, E. and GABRIEL, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63** 655-660.

MAURER, W., HOTHORN, L. A. and LEHMACHER, W. (1995). Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses. In *Biometrie in der chemische-pharmazeutichen Industrie* (J. VOLLMAN, ed.) **6**. Fischer Verlag, Stuttgart.

ROMANO, J. P., SHAIKH, A. and WOLF, M. (2011). Consonance and the closure method in multiple testing. *The International Journal of Biostatistics* **7**.

ROMANO, J. P. and WOLF, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *J. Am. Statist. Assoc.* **100** 94-108.

ROSENBAUM, P. R. (2008). Testing hypotheses in order. *Biometrika* **95** 248-252.

ROSENBLUM, M. and VAN DER LAAN, M. J. (2011). Optimizing randomized trial designs to distinguish which subpopulations benefit from treatment. *Biometrika* **98** 845-860.

SLAMON, D. J., LEYLAND-JONES, B., SHAK, S., FUCHS, H., PATON, V., BAJAMONDE, A., FLEMING, T., EIERMANN, W., WOLTER, J., PEGRAM, M., BASELGA, J. and NORTON, L. (2001). Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *New England Journal of Medicine* **344** 783-792.

SONG, Y. and CHI, G. Y. H. (2007). A method for testing a prespecified subgroup in clinical trials. *Statistics in Medicine* **26** 3535–3549.

SONNEMANN, E. and FINNER, H. (1988). Vollständigkeitssätze für multiple Testprobleme. In *Multiple Hypothesenprüfung* (P. Bauer, G. Hommel and E. Sonnemann, eds.) 121-135. Springer, Berlin.

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

BALTIMORE, MD, 21205, USA
E-MAIL: mrosenbl@jhsph.edu